

Data collection and sampling

6

syllabus reference

Data analysis 2

- Data collection and sampling

In this chapter

- 6A Target populations and sampling
- 6B Population characteristics
- 6C Bias
- 6D Types of data
- 6E Estimating populations



are you **READY?**



Are you ready?

Try the questions below. If you have difficulty with any of them, extra help can be obtained by completing the matching **SkillSHEET**. Either click on the **SkillSHEET** icon next to the question on the *Maths Quest Preliminary Course CD-ROM* or ask your teacher for a copy.



Writing one amount as a percentage of another

- 1 In each of the following write the first quantity as a percentage of the second.
- a** 12; 200 **b** 54; 250 **c** 45; 527



Unitary method of percentages

- 2 Find the total quantity given that:
- a** 5% of it is 230 **b** 15% of it is 126 **c** 2.4% of it is 28.8.



Finding a percentage of a quantity

- 3 Find:
- a** 15% of 900 **b** 6% of 1150 **c** 8.25% of 1327.

Target populations and sampling

In the previous chapter we looked at statistics and the role they play in society. In this chapter we investigate in more detail how data are collected.

To gather data we conduct a **poll**. A poll is the recording of responses to a set of questions known as a **questionnaire**.

Gallup poll

The most famous poll is named after its founder, the American statistician, George Gallup, who was born in 1901.

Find out about Gallup and his work and how Gallup polls are used today.

The first step in gathering the relevant data for a statistical investigation is to target the **population** to be investigated. This means identifying the sections of the population for whom the statistical investigation will have meaning.

For example, if investigating the medical needs of a community, we would not conduct our survey at the local fitness club. For such a survey, we would survey doctors and other medical personnel, as well as a selection of patients who use the existing facilities.

Identifying the target population

For each of the following statistical investigations, identify the population that you would target for a survey.

- 1 The school 'End of Year' Committee wants to find out the preferred venue, band and meals for the Year 12 farewell.
- 2 The local council wants to know what sporting facilities are needed in the local area.
- 3 A newspaper wants a survey to predict the winner of a forthcoming election.
- 4 A group of people planning to build a preschool would like to know what facilities attract people to a particular preschool.
- 5 A recording label wants to estimate the potential success of a 'grunge' band.

When starting an investigation, we must determine the quantity of data needed for the **database**. Consider the case of a company calculating the TV ratings. Does the company need to find out what every household is watching? Obviously they do not; they ask a selection of homes to record their TV viewing.

Now consider the case of selecting a band to play at the Year 12 farewell. In this case it is reasonable to ask every Year 12 student their opinion.

Data can be collected in one of two ways:

1. **Census**. In a census an entire population is counted. Australians complete 'The Census' every five years. This is a survey of every household in the nation conducted by the Bureau of Statistics. For the purposes of most statistical investigations, a census surveys everyone in the target population, such as the Year 12 example above.
2. **Sample**. A sample is a more practical method for doing most surveys. Only a selection of the target population is surveyed, with the results taken to be representative of the whole group. The TV ratings example is one where a sample is used.

WORKED Example 1

In each of the following, state if the information was obtained by census or sample.

- a** A school uses the roll to count the number of students absent each day.
- b** The television ratings, in which 2000 families complete a survey on what they watch over a one week period.
- c** A light globe manufacturer tests every hundredth light globe off the production line.
- d** A teacher records the examination marks of her class.

THINK

- a** Every student is counted at roll call each morning.
- b** Not every family is asked to complete a ratings survey.
- c** Not every light globe is tested.
- d** The marks of every student are recorded.

WRITE

- a** Census
- b** Sample
- c** Sample
- d** Census

To ensure that the results of your sample are representative of the whole population, the method of sampling is important. There are three main methods of choosing a sample: random sample, stratified sample and systematic sample.

Method 1. Random sample. In a random sample those to be surveyed are selected by chance. When a random sample is conducted, every person in the target population should have an equal chance of being selected. For example, the names of the people to complete your survey may be drawn from a hat. If this method is used, you should get a good mixture of people in your survey.

Suppose that we are going to survey students in a school. We want a mixture of students and could choose a fixed number of students from each year. Suppose we decide to survey 60 students. We could select 10 from each year, but if we did this the survey would not have the correct proportion of students from each year. For example, 20% of the students at this school are in Year 7, but only $16\frac{2}{3}\%$ of the survey participants are in Year 7.

If we randomly select the 60 participants, we could expect those chosen to resemble the overall population. To get a genuinely random sample, you could use the random number generator on your calculator.

The calculator will give you a random decimal number between 0 and 1. Each random number is multiplied by 1000 and then rounded up to give a whole number between 1 and 1000. We could then take the roll number of the student corresponding to this number. Repeating this 60 times will give us a genuinely random sample.

Year	No. of students
Year 7	200
Year 8	180
Year 9	190
Year 10	185
Year 11	135
Year 12	110
Total	1000

WORKED Example 2

Three students from a school are to be selected to participate in a statewide survey of school students. There are 750 students at the school. To choose the participants, a random number generator is used with the results 0.983, 0.911 and 0.421. What are the roll numbers of the students who should be selected?

THINK

- 1 Multiply the results of the random number generator by the size of the population.
- 2 Round up to whole numbers.

WRITE

$$0.983 \times 750 = 737.25$$

$$0.911 \times 750 = 683.25$$

$$0.421 \times 750 = 315.75$$

The 738th, 684th and the 316th people on the roll would be surveyed.

Any other method may not give a truly representative sample. For example, if you survey people in the playground you may:

- have a tendency to ask people you know
- choose an area where a lot of students from a particular year tend to sit
- choose more of one sex than the other.

Method 2. Stratified sample. In this type of sample you deliberately choose people to complete your survey who are representative of the whole population. In the school survey you would need to select six **strata** that had the correct proportion of students from each year. In this example, 20% of the school population are in Year 7 and so 20% of your sample should be from Year 7.

WORKED Example 3

Adrian is conducting a survey of school students. At his school, 47% of the population are male and 53% are female. If Adrian decides to survey 60 students, how many students of each sex should he choose if he decides to use a stratified sample?

THINK

- 1 The population is 47% male and so 47% of the sample should be male. Find 47% of 60.
- 2 The population is 53% female and so 53% of the sample should be female. Find 53% of 60.
- 3 Make a conclusion about how many of each sex should participate in the survey, rounding off your answers.

WRITE

$$47\% \text{ of } 60 = 0.47 \times 60 \\ = 28.2$$

$$53\% \text{ of } 60 = 0.53 \times 60 \\ = 31.8$$

There should be 28 males and 32 females in Adrian's survey.

Method 3. Systematic sample. Systematic sampling is where those chosen for the sample are chosen in a systematic or organised way. This method is mostly used in quality control situations discussed in the previous chapter. For example, suppose that the quality and strength of sports shoes is being tested. The quality control department may test every 20th pair of shoes that come off the production line. In doing a survey, every 20th person on the school roll may be surveyed.

remember

1. Before beginning a statistical investigation it is important to identify the target population.
2. The survey can be conducted either by:
 - (a) Census — the entire target population is surveyed, or
 - (b) Sample — a selection is surveyed such that those selected are representative of the entire target population.
3. There are three methods for selecting a sample.

Method 1. Random sample — chance is the only factor in deciding who is surveyed. This is best done using a random number generator.

Method 2. Stratified sample — those sampled are chosen in proportion to the entire population.

Method 3. Systematic sample — a system is used to choose those who are to be in the sample.

EXERCISE 6A

Target populations and sampling

WORKED Example 1

- 1 A school conducts an election for a new school captain. Every teacher and student in the school votes. Is this an example of a census or a sample? Explain your answer.
- 2 A survey is conducted by a council to see what sporting facilities the community needs. If 500 people who live in the community are surveyed, is this an example of a census or a sample?
- 3 For each of the following surveys, state whether a census or a sample has been used.
 - a Two hundred people in a shopping centre are asked to nominate the supermarket where they do most of their grocery shopping.
 - b To find the most popular new car on the road, 500 new car buyers are asked what make and model car they purchased.
 - c To find the most popular new car on the road, the make and model of every new car registered are recorded.
 - d To find the average mark in the Mathematics half-yearly exam, every student's mark is recorded.
 - e To test the quality of tyres on a production line, every 100th tyre is road tested.
- 4 For each of the following, recommend whether you would use a census or a sample to obtain the results.
 - a To find the most watched television program on Monday night at 7:30 pm
 - b To find the number of cars sold during a period of one year
 - c To find the number of cars that pass through the tollgates on the Sydney Harbour Bridge each day
 - d To find the percentage of computers produced by a company that are defective
- 5 An opinion poll is conducted to try to predict the outcome of an election. Two thousand people are telephoned and asked about their voting intention. Is this an example of a census or a sample?

**WORKED
Example****2**

- 6 A factory has 500 employees. Each employee has an employee number between 1 and 500. Five employees are selected to participate in an Occupational Health and Safety survey. To choose the participants, a random number generator is used. The results are 0.326, 0.352, 0.762, 0.989 and 0.018. What are the employee numbers of those to participate in the survey?
- 7 A school has 837 students. A survey of 10 students in the school is to be conducted. A random number generator is used to select the participants. If the random numbers chosen are:
0.988 0.251 0.498 0.661 0.247 0.031 0.967 0.932 0.229 0.443
what are the roll numbers of the students who should be selected?
- 8 A survey is to be conducted of 20 out of 50 000 people in a country town. Those selected are to be chosen using a random number generator.
- a Use your calculator to generate 20 random numbers.
 - b Calculate the electoral roll numbers of the people who should be chosen for the survey.
- 9 For each of the following, state whether the sample used is an example of random, stratified or systematic sampling.
- a Every 10th tyre coming off a production line is tested for quality.
 - b A company employs 300 men and 450 women. The sample of employees chosen for a survey contains 20 men and 30 women.
 - c The police breathalyse the driver of every red car.
 - d The names of the participants in a survey are drawn from a hat.
 - e Fans at a football match fill in a questionnaire. The ground contains 8000 grandstand seats and 20 000 general admission seats. The questionnaire is then given to 40 people in the grandstand and 100 people who paid for a general admission seat.

10 multiple choice

Which of the following is an example of a systematic sample?

- A The first 20 students who arrive at school each day participate in the survey.
- B Twenty students to participate in the survey are chosen by a random number generator.
- C Twenty students to participate in the survey are selected in proportion to the number of students in each school year.
- D Ten boys and 10 girls are chosen to participate in the survey.

11 multiple choice

Which of the following statistical investigations would be practical to complete by census?

- A A newspaper wants to know public opinion on a political issue.
- B A local council wants to know if a skateboard ramp would be popular with young people in the area.
- C An author wants a cricket player's statistics for a book being written.
- D An advertising agency wants to know the most watched program on television.

**WORKED
Example****3**

- 12 Zara is conducting a survey of the people at work. At her work 68% of the workers are male and 32% are female. If Zara decides to survey 50 workers, how many of each sex should she choose, if she chooses to use a stratified sample?

- 13** The table below shows the percentage of students that are in each year of a school.

Year	Percentage of students
7	20%
8	19%
9	21%
10	16%
11	13%
12	11%

If 40 students are to participate in a survey, find the number chosen from each year group if stratified sampling is used.

Census or sample

For each of the following statistical investigations, state whether you would gather data using a census or sample. For those for which you would use a sample, state the best method for selecting the sample.

- 1** A company wants to test the life of its batteries.
- 2** A sporting club wants to elect a new club president.
- 3** A market research company wants to determine the most popular brand of toothpaste.
- 4** A theme park wants to know from which state and suburb its visitors come.
- 5** A Gallup poll is conducted to determine the preferred prime minister.

Population characteristics

Characteristics about a population can be estimated by taking a sample of that population. For example, to estimate the average income of Australians we could conduct an anonymous survey of a sample of the Australian population.

To get an accurate estimate, it is important that the sample taken has similar characteristics to the entire population. For example, the sample should have an equal number of males and females, and the ages of those in the sample should be in the same proportion as for the whole population. The area sampled should include a cross-section of people according to socio-economic level, ethnic background, religion etc., similar to that of the whole population. In such cases we use the stratified sampling technique.

Characteristics (such as sex and age) of the population and the sample must also match, if we expect our sample to have incomes in the same proportions as those of the general population.

Population characteristics

If we are trying to estimate the average income of Australian families, what would be the effect if our sample:

- 1 contained a majority of one sex?
- 2 contained only people in the 15–21 age group?
- 3 contained only people who live in a small country town?

To calculate the number of participants in a sample from each strata of the population, we calculate each strata as a fraction of the total population and multiply this fraction by the total number to be chosen in the sample.

WORKED example 4

The table at right shows the enrolment at a high school.

Sam is conducting a music survey for the school disco. He plans to survey 60 students. Show how Sam should break down his sample in terms of year and sex.

Year	Boys	Girls
7	96	102
8	85	87
9	92	88
10	80	74
11	71	75
12	69	65
Total	493	491
Grand total	984	

THINK

- 1 Write each strata as a fraction of the total and then multiply by the 60 to be chosen in the sample.

WRITE

$$\begin{aligned}\text{Year 7 boys} &= \frac{96}{984} \times 60 \\ &\approx 5.85\end{aligned}$$

$$\begin{aligned}\text{Year 7 girls} &= \frac{102}{984} \times 60 \\ &\approx 6.22\end{aligned}$$

$$\begin{aligned}\text{Year 8 boys} &= \frac{85}{984} \times 60 \\ &\approx 5.18\end{aligned}$$

$$\begin{aligned}\text{Year 8 girls} &= \frac{87}{984} \times 60 \\ &\approx 5.30\end{aligned}$$

$$\begin{aligned}\text{Year 9 boys} &= \frac{92}{984} \times 60 \\ &\approx 5.61\end{aligned}$$

$$\begin{aligned}\text{Year 9 girls} &= \frac{88}{984} \times 60 \\ &\approx 5.37\end{aligned}$$

$$\begin{aligned}\text{Year 10 boys} &= \frac{80}{984} \times 60 \\ &\approx 4.88\end{aligned}$$

Continued over page 

THINK

- 2 To complete the table, round off each of these answers to the nearest whole number.

WRITE

$$\begin{aligned}\text{Year 10 girls} &= \frac{74}{984} \times 60 \\ &\approx 4.51\end{aligned}$$

$$\begin{aligned}\text{Year 11 boys} &= \frac{71}{984} \times 60 \\ &\approx 4.33\end{aligned}$$

$$\begin{aligned}\text{Year 11 girls} &= \frac{75}{984} \times 60 \\ &\approx 4.57\end{aligned}$$

$$\begin{aligned}\text{Year 12 boys} &= \frac{69}{984} \times 60 \\ &\approx 4.21\end{aligned}$$

$$\begin{aligned}\text{Year 12 girls} &= \frac{65}{984} \times 60 \\ &\approx 3.96\end{aligned}$$

Number of students to be sampled

Year	Boys	Girls
7	6	6
8	5	5
9	6	5
10	5	5
11	4	5
12	4	4
Total	30	30
Grand total	60	

Note that on some occasions after rounding off each of the answers, the total number of people to participate in the sample may add to one more than the number that we planned to select. In such cases, the person doing the sample should include this extra person as it gives a better sample of the overall population.

remember

1. Stratified sampling is used to ensure that a sample chosen is representative of the entire population.
2. To choose the number of participants from each stratum in the population, write each stratum as a fraction of the total population, and then multiply by the number of people you intend to select in the sample. Round off each of the answers to this multiplication to the nearest whole number.
3. Sometimes, as a result of rounding off, there may be one more to be chosen in the sample than was originally intended.

EXERCISE 6B

Population characteristics

**WORKED
Example**

4

- 1 The table below shows the number of students in each year at a school.

Year	No. of students
7	90
8	110
9	90
10	80
11	70
12	60
Total	500

If a survey is to be given to 50 students at the school, how many from each year should be chosen if a stratified sample is used?



6.1 SkillSheet
Writing one amount as a percentage of another

6.2 SkillSheet
Unitary method of percentages

6.3 SkillSheet
Finding a percentage of a quantity

- 2 A company employs 300 men and 200 women. If a survey of 60 employees using a stratified sample is completed, how many people of each sex participated?
- 3 A business has 400 employees of which 250 are female and 150 are male. The business intends to survey 40 of their employees. If a stratified survey is to be conducted, how many employees of each sex should be surveyed?
- 4 In the head office of a bank there are 250 employees. Ten of these employees are senior management, 60 are middle management and 180 are employed as clerks. A survey is to be conducted of 50 staff members. How many employees at each level should be surveyed?
- 5 The Department of Education wants to survey a school population. At the school there are 93 teachers and 1248 students. If the department is to survey a total of 50 people, how many teachers and how many students should participate in the survey?
- 6 The table below shows the age and sex of the staff of a corporation.

Age	Male	Female
20–29	61	44
30–39	40	50
40–49	74	16
50–59	5	10

A survey of 50 employees is to be done. Using a stratified survey, suggest the breakdown of people to participate in terms of age and sex.

- 7 The table below shows the number of students who are in each year level at a school.

Year	Number of students
7	187
8	192
9	168
10	157
11	137
12	108

If 80 students are to be selected to participate in a survey, how many should be chosen from each year level?

- 8 A shopping centre has a floor area of 5000 m^2 . There is one major department store with an area of 1500 m^2 , two smaller department stores of 750 m^2 and 40 small stores of 50 m^2 . The management of the centre assigns voting rights in the shopping centre in proportion to the floor area of each business. Given that there are 200 votes to be distributed, how many votes should each business get?
- 9 The table below shows the population of a school.

Year	Boys	Girls
7	104	98
8	112	119
9	107	110
10	97	88
11	75	82
12	68	66

A survey of 100 students is to be conducted. Complete the table below to show the number of students from each year and sex who should participate in a stratified sample.

Year	Boys	Girls
7		
8		
9		
10		
11		
12		

Choosing a sample

Consider how you would choose your sample if you wished to conduct a survey for your next school disco. Use the method in worked example 4 to select the number of boys and girls that should be chosen from each year to do your survey.

Step 1. Find out the number of boys and girls enrolled in each year at your school.

Step 2. Calculate the percentage of the whole school population in each year for both boys and girls.

Step 3. Choose a suitable sample size and calculate the number of boys and girls needed from each year to complete your survey.

10 QUICK QUESTIONS 1

For each of the following, state whether a census or sample has been used.

- 1 A school votes to elect a school captain.
- 2 Five hundred drivers complete a survey on the state of a major highway.
- 3 Every insurance customer completes a questionnaire when renewing their policy.

For each of the following, state the type of sample that has been taken.

- 4 A computer selects 500 phone numbers.
- 5 Every 1000th person in the telephone book.
- 6 Private and business telephone numbers are chosen in proportion to the number of private and business listings.
- 7 In a school there are 1000 students and 100 teachers. A sample of 20 members of the school is to participate in an Occupational Health and Safety review. If a stratified sample is chosen, how many students should participate?
- 8 In the Parliament there are 90 Liberals and 60 Labor members. For a committee, 3 Liberal and 3 Labor members are selected. Explain why this is not a stratified sample.
- 9 For question 8, calculate the correct number of people who should have been selected from each party.
- 10 Give an advantage that stratified sampling has over random sampling.

Bias

No doubt you have heard the comment, ‘There are lies, damned lies and statistics’. This implies that we should be wary of statistical figures quoted. Indeed, we should always make informed decisions of our own and not simply accept the mass of statistics that bombards us through the media.

Bias can be introduced into statistics by:

1. questionnaire design
2. sampling bias
3. the interpretation of results.

Bias in questionnaire design

Consider a survey designed to collect data on opinions relating to culling kangaroo numbers in Australia.

The questions may be designed to be *emotive* in nature. Respondents in these situations feel obliged to show compassion. Posing a question in the form, ‘The kangaroo is identified as a native Australian animal, not found anywhere else in the world. Would you be in favour of culling kangaroos in Australia?’, would almost certainly encourage a negative response.

Using a *leading* question (one which leads the respondent to answer in a particular way) can cause bias to creep into responses. Rephrasing the question in the form, ‘As you know, kangaroos cause massive damage on many farming properties. You’d agree that their numbers need culling, wouldn’t you?’, would encourage a positive response.

Using *terminology* that is unfamiliar to a large proportion of those being surveyed would certainly produce unreliable responses. ‘Do you think we need to cull herbivorous marsupial mammals in Australia?’, would cause most respondents to answer according to their understanding of the terms used. If the survey was conducted by an interviewer, the term could be explained. In the case of a self-administered survey, there would be no indication of whether the question was understood or not.



Sampling bias

As discussed previously, an ideal sample should reflect the characteristics of the population. Statistical calculations performed on the sample would then be a reliable indication of the population’s features.

Selecting a sample using a *non-random method*, as discussed earlier, generally tends to introduce an element of bias.

Particular responses can be selected from all those received. In collecting information on a local issue, an interviewer on a street corner may record responses from many passers-by. From all the data collected, a sample could be chosen to support the issue, or alternatively another sample could be chosen to refute the same issue.

A sample may be selected under *abnormal conditions*. Consider a survey to determine which lemonade was more popular — Kirks or Schweppes. Collecting data one week when one of the brands was on special at half price would certainly produce misleading results.

Data are often collected by radio and television stations via *telephone polls*. A ‘Yes’ response is recorded on a given phone-in number, while the ‘No’ respondents are asked to ring a different phone-in number. This type of sampling does not produce a representative sample of the population. Only those who are highly motivated tend to ring and there is no monitoring of the number of times a person might call, recording multiple votes.

When data are collected from mailing surveys, bias results if the *non-response rate* is high (even if the selected sample was a random one). The responses received often represent only those with strong views on the subject, while those with more moderate views tend to lack representation in their correct proportion.

Statistical interpretation bias

Once the data have been collected, collated and subjected to statistical calculations, bias may still occur in the interpretation of the results.

Misleading graphs can be drawn leading to a biased interpretation of the data. Graphical representations of a set of data can give a visual impression of ‘little change’ or ‘major change’ depending on the scales used on the axes.

The use of terms such as ‘majority’, ‘almost all’ and ‘most’ are open to interpretation. When we consider that 50.1% ‘for’ and 49.9% ‘against’ represents a ‘majority for’ an issue, the true figures have been hidden behind words with very broad meanings. Although we would probably not learn the real facts, we should be wary of statistical issues quoted in such terms.

Bias in statistics

The aim of this investigation is to study statistical data that you suspect to be biased.

Conduct a search of newspapers, magazines or any printed material to collect instances of quoted statistics that you believe to be biased. There are occasions when television advertisements quote statistical figures as a result of questionable sampling techniques. For each example, discuss:

- 1 the purpose of the survey
- 2 how the data might have been collected
- 3 the question(s) that may have been asked (try to pose the question(s) in a variety of ways to influence different outcomes)
- 4 ways in which bias might be introduced
- 5 variations in interpretation of the data.

Biased sampling

Discuss the problems that would be caused by each of the following biased samples.

- 1 A survey is to be conducted to decide the most popular sport in a local community. A sample of 100 people was questioned at a local football match.
- 2 A music store situated in a shopping centre wants to know the type of music that it should stock. A sample of 100 people was surveyed. The sample was taken from people who passed by the store between 10 and 11 am on a Tuesday.
- 3 A newspaper conducting a Gallup poll on an election took a sample of 1000 people from Sydney.

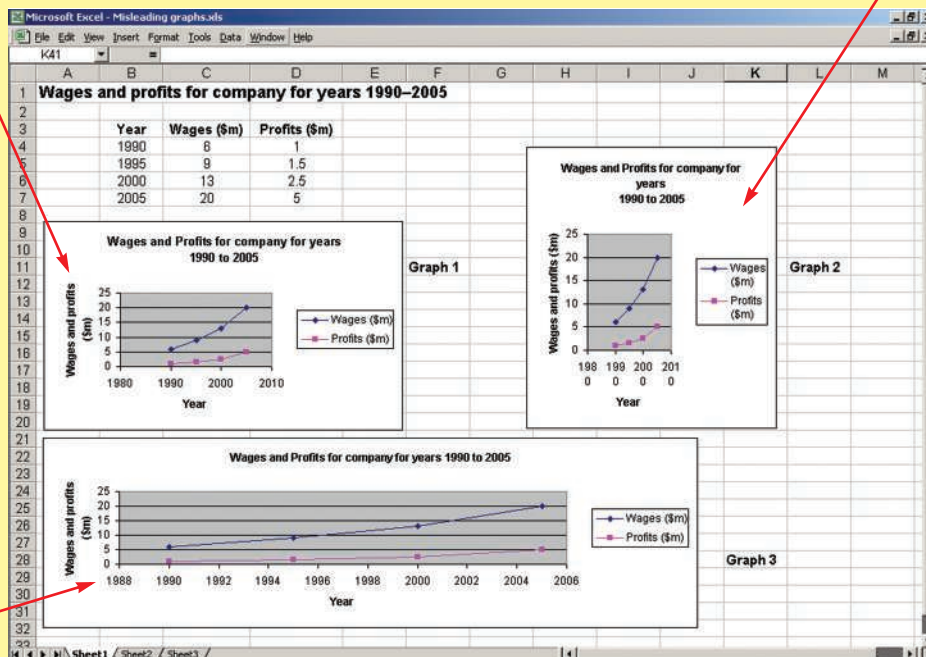
Spreadsheets creating misleading graphs

Let us practise producing misleading graphs. Consider the data in this table.

Year	1990	1995	2000	2005
Wages (\$m)	6	9	13	20
% increase in wages	25	50	44	54
Profits (\$m)	1	1.5	2.5	5
% increase in profits	20	50	66	100

We shall use a spreadsheet to produce misleading graphs based on these data.

Graph 1



Graph 2

Graph 3

- 1 Enter the data as indicated in the spreadsheet (see page 182).
- 2 Graph the data using the **Chart Wizard**. You should obtain a graph similar to Graph 1.
- 3 Copy and paste the graph twice within the spreadsheet.
- 4 Graph 2 gives the impression that the wages are a great deal higher than the profits. This effect was obtained by reducing the horizontal axis. Experiment with shortening the horizontal length and lengthening the vertical axis.
- 5 In Graph 3 we get the impression that the wages and profits are not very different. This effect was obtained by lengthening the horizontal axis and shortening the vertical axis. Experiment with various combinations.
- 6 Print out your three graphs and examine their differences.

Note that all three graphs have been drawn from the same data using valid scales. A cursory glance leaves us with three different impressions. Clearly, it is important to look carefully at the scales on the axes of graphs.

Another method which could be used to change the shape of a graph is to change the scale of the axes.

- 7 Right click on the axis value, enter the **Format axis** option, click on the **Scale** tab, then experiment with changing the scale values on both axes.

Techniques such as these are used to create different visual impressions of the same data.

- 8 Use the data in the table to create a spreadsheet, then produce two graphs depicting the percentage increase in both wages and profits over the years giving the impression that:
 - a the profits of the company have not grown at the expense of wage increases (the percentage increase in wages is similar to the percentage increase in profits)
 - b the company appears to be exploiting its employees (the percentage increase in profits is greater than that for wages).

WORKED Example 5

Discuss why the following selected samples could provide bias in the statistics collected.

- a In order to determine the extent of unemployment in a community, a committee phoned two households (randomly selected) from each page of the local telephone book during the day.
- b A newspaper ran a feature article on the use of animals to test cosmetics. A form beneath the article invited responses to the article.

THINK

- a 1 Consider phone book selection.
- 2 Consider those with no phone contact.
- 3 Consider the hours of contact.

WRITE

- a Phoning two randomly selected households per page of the telephone directory is possibly a representative sample.
However, those without a home phone and those with unlisted numbers could not form part of the sample.
An unanswered call during the day would not necessarily imply that the resident was at work.

Continued over page 

THINK

- b** ① Consider the newspaper circulation.
- ② Consider the urge to respond.

WRITE

- b** Selecting a sample from a circulated newspaper excludes those who do not have access to the paper.
In emotive issues such as these, only those with strong views will bother to respond, so the sample will represent extreme points of view.

remember

Bias can be introduced at each of the following stages:

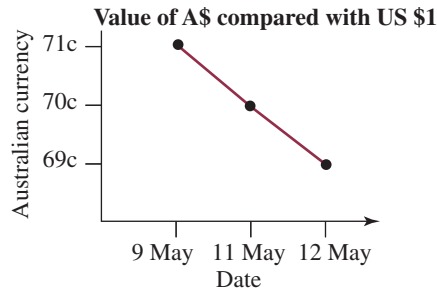
1. questionnaire design
2. sampling bias
3. interpretation of results.

EXERCISE 6C**Bias**

- 1 Rewrite the following questions, removing any elements or words that might contribute to bias in responses.
 - a** The poor homeless people, through no fault of their own, experience great hardship during the freezing winter months. Would you contribute to a fund to build a shelter to house our homeless?
 - b** Most people think that, since we've developed as a nation in our own right and broken many ties with Great Britain, we should adopt our own national flag. You'd agree with this, wouldn't you?
 - c** You'd know that our Australian 50 cent coin is in the shape of a dodecagon, wouldn't you?
 - d** Many in the workforce toil long hours for low wages. By comparison, politicians seem to get life pretty easy when you take into account that they only work for part of the year and they receive all those perks and allowances. You'd agree, wouldn't you?
- 2 Rewrite parts **a** to **d** in question 1 so that the expected response is reversed.
- 3 What forms of sampling bias can you identify in the following samples?
 - a** Choosing a sample from students on a bus travelling to a sporting venue to answer a questionnaire regarding sporting facilities at their school
 - b** Sampling using 'phone-in' responses to an issue viewed on a television program
 - c** Promoting the results of a mail-response survey when fewer than half the selected sample replied
 - d** Comparing the popularity of particular chocolate brands when one brand has a 'two for the price of one' special offer
 - e** Choosing a Year 7 class and a Year 12 class to gather data relating to the use of the athletics oval after school



4 Why does this graph produce a biased visual impression?



5 Comment on the following statement:

‘University tests have demonstrated that *Double-White* toothpaste is consistently used by the majority of teenagers and is more effective than most other toothpastes.’

6 Surveys are conducted on samples to determine the characteristics of the population. Discuss whether the samples selected would provide a reliable indication of the population’s characteristics.

Sample	Population
a Year 11 students	Student drivers
b Year 12 students	Students with part-time jobs
c Residents attending a neighbourhood watch meeting	Residents of a suburb
d Students in the school choir	Music students in the school
e Cars in a shopping centre car park	Models of Holden cars on the road
f Males at a football match	Popular TV programs
g Users of the local library	Popular teenage magazines

Bias

It is important that a sample is chosen randomly to avoid bias.

Consider the following situation.

The government wants to improve sporting facilities in Sydney. They decide to survey 1000 people about what facilities they would like to see improved. To do this, they choose the first 1000 people through the gate at a football match at the Sydney Cricket Ground.

In this situation it is likely that the results will be biased towards improving facilities for football. It is also unlikely that the survey will be representative of the whole population in terms of equality between men and women, age of the participants and ethnic backgrounds.

Questions can also create bias. Consider asking the question, ‘Is football your favourite sport?’ The question invites the response that football is the favourite sport rather than allowing a free choice from a variety of sports by the respondent.

Consider each of the following surveys and discuss:

- a any advantages, disadvantages and possible causes of bias
- b a way in which a truly representative sample could be obtained.

- 1 Surveying food product choices by interviewing customers of a large supermarket chain as they emerge from the store between 9.00 am and 2.00 pm on a Wednesday.
- 2 Researching the popularity of a government decision by stopping people at random in a central city mall.
- 3 Using a telephone survey of 500 people selected at random from the phone book to find if all Australian States should have Daylight Saving Time in summer.
- 4 A bookseller uses a public library database to survey for the most popular novels over the last three months.
- 5 An interview survey about violence in sport taken at a rugby league football venue as spectators leave.



Types of data

Data can be put into two categories.

1 — Categorical data

Categorical data cannot be measured; they can only be put into categories.

An example of categorical data is makes of cars. The categories for the data would be all possible makes of cars such as Ford, Holden, Toyota, Mazda etc. Other questions that would lead to categorical data would be things such as:

- What is your hair colour?
- Who is your favourite musical performer?
- What method of transport do you use to get to school?

2 — Quantitative data

Quantitative data can be measured. They are data to which we can assign a numerical value.

Quantitative data are collected either by measurement or by counting. For example, the data collected by measuring the heights of students are quantitative data. The data collected by counting the ages of students in years are also quantitative data.

WORKED Example 6

State whether the following pieces of data are categorical or quantitative.

- a** The value of sales recorded at each branch of a fast-food outlet
- b** The breeds of dog that appear at a dog show

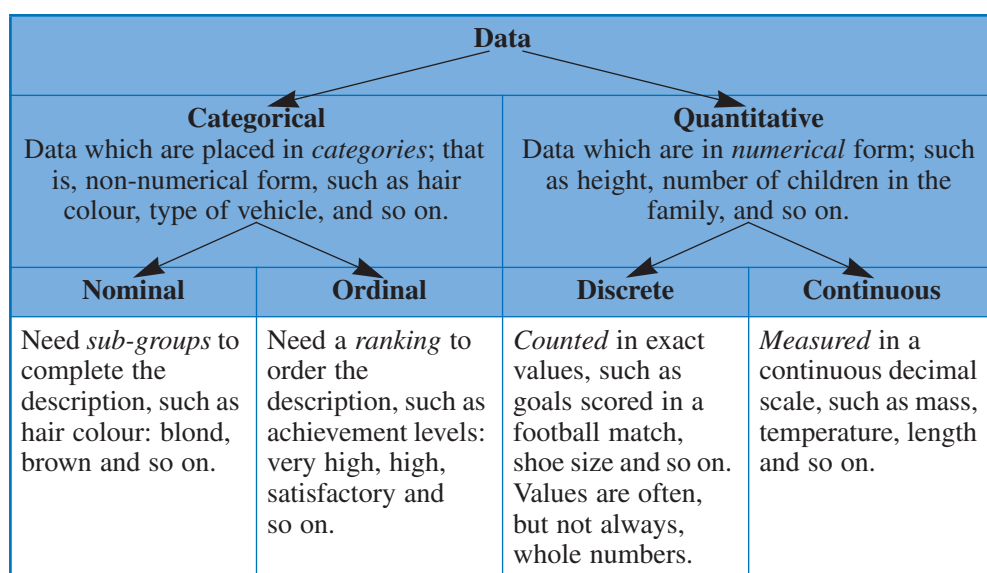
THINK

- a** The value of sales at each branch can be measured.
- b** The breeds of dog at a show cannot be measured.

WRITE

- a** The value of sales are quantitative data.
- b** The breeds of dog are categorical data.

There are two types of categorical data and two types of quantitative data.



WORKED Example 7

Classify each of the following data using two selections from the following descriptive words: categorical, quantitative, nominal, ordinal, discrete and continuous.

- a** the number of students absent from school
- b** the types of vehicle using a certain road
- c** the various pizza sizes available at a local takeaway
- d** the room temperature at various times during a particular day

THINK

- a** ① Determine whether the data are categorical or quantitative.
- ② Determine whether the data are discrete or continuous.

WRITE

- a** The data are quantitative as absences are represented by a number.
The data are discrete as the number of absences can be counted and is an exact value.

Continued over page 

THINK

- b** ① Determine whether the data are categorical or quantitative.
 ② Determine whether the data are nominal or ordinal.
- c** ① Determine whether the data are categorical or quantitative.
 ② Determine whether the data are nominal or ordinal.
- d** ① Determine whether the data are categorical or quantitative.
 ② Determine whether the data are discrete or continuous.

WRITE

- b** The data are categorical as the types of vehicle need to be placed in non-numerical categories. The data are nominal as there is no ranking or order involved.
- c** The data are categorical as the pizza sizes need to be ranked in order ranging from small to family. The data are ordinal as pizzas are ranked in order of size.
- d** The data are quantitative as room temperature is represented by a number. The data are continuous as temperature can assume any value and measurement is involved.

remember

1. Data can be classified as either:
 - (a) categorical — the data are in categories, or
 - (b) quantitative — the data can be either measured or counted.
2. Categorical data can be either:
 - (a) nominal — where the order of the categories is not important, or
 - (b) ordinal — the categories have a logical order.
3. Quantitative data can be either:
 - (a) discrete — the data can take only certain values, usually whole numbers, or
 - (b) continuous — the data can take any value depending on the degree of accuracy.

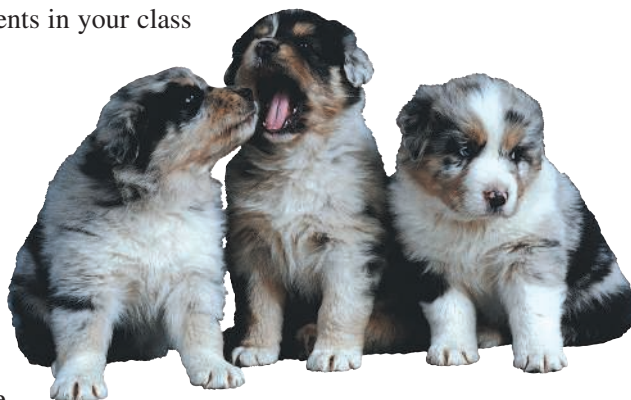
EXERCISE**6D****Types of data****WORKED
Example****6**

- 1 State whether the data collected in each of the following situations would be categorical or quantitative.
- a** The number of matches in each box is counted for a large sample of boxes.
 - b** The sex of respondents to a questionnaire is recorded as either M or F.
 - c** A fisheries inspector records the lengths of 40 cod.
 - d** The occurrence of hot, warm, mild and cool weather for each day in January is recorded.
 - e** The actual temperature for each day in January is recorded.
 - f** Cinema critics are asked to judge a film by awarding it a rating from one to five stars.

- 2** State whether the categorical data formed by each of the following situations are nominal or ordinal.
- a** On a school report students are ranked as Unsatisfactory, Satisfactory, Above average or Outstanding.
 - b** The day of the week that a business has the most customers is recorded.
 - c** Visitors to a museum are recorded as being either male or female.
 - d** The colour of each traffic light on a journey is recorded.
 - e** The make of each television in an electronics store is recorded.
- 3** State whether the quantitative data formed by each of the following situations are discrete or continuous.
- a** The heights of 60 tomato plants at a plant nursery
 - b** The number of jelly beans in each of 50 packets
 - c** The time taken for each student in a class of six-year-olds to tie their shoelaces
 - d** The petrol consumption rate of a large sample of cars
 - e** The IQ (intelligence quotient) of each student in a class
- 4** Classify each of the following data using two words selected from the following descriptive words: categorical, quantitative, nominal, ordinal, discrete and continuous.
- a** The population of your town or city
 - b** The types of motorbike in a parking lot
 - c** The heights of people in an identification line-up
 - d** The masses of babies in a group
 - e** The languages spoken at home by students in your class
 - f** The time spent watching TV
 - g** The number of children in the families in your suburb
 - h** The air pressure in your car's tyres
 - i** The number of puppies in a litter
 - j** The types of radio program listened to by teenagers
 - k** The times for swimming 50 metres
 - l** The quantity of fish caught in a net
 - m** The number of CDs you own
 - n** The types of shops in a shopping centre
 - o** The football competition ladder at the end of each round
 - p** The lifetime of torch batteries
 - q** The number of people attending a rock concert
 - r** Exam grades
 - s** The types of magazine sold at a newsagency
 - t** Hotel accommodation rating
- 5** For each of the following, state if the data are categorical or quantitative. If quantitative, state if the data are discrete or continuous.
- a** The number of students in each class at your school
 - b** The teams people support at a football match
 - c** The brands of peanut butter sold at a supermarket
 - d** The heights of people in your class
 - e** The interest rate charged by each bank
 - f** A person's pulse rate

**WORKED
Example**

7



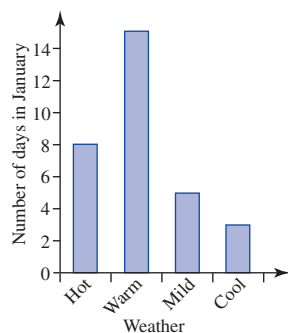
- 6 An opinion poll was conducted. A thousand people were given the statement 'Euthanasia should be legalised'. Each person was offered five responses: strongly agree, agree, unsure, disagree and strongly disagree. Describe the data type in this example.
- 7 A teacher marks her students' work with a grade A, B, C, D, or E. Describe the data type used.
- 8 A teacher marks his students' work using a mark out of 100. Describe the data type used.

9 **multiple choice**

The number of people who are using a particular bus service are counted over a two week period. The data formed by this survey would be an example of:

- A categorical and ordinal data
 B cateorical and nominal data
 C quantitative and discrete data
 D quantitative and continuous data

- 10 The following graph shows the number of days of each weather type for the Gold Coast in January.

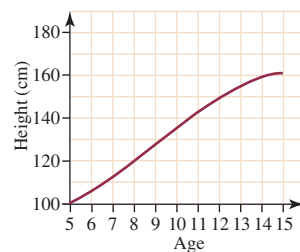


Describe the data in this example.



- 11 The graph at right shows a girl's height each year for 10 years.

Describe the data in this example.



10 QUICK QUESTIONS 2

- 1 'The rich should pay more in tax, to allow the poorer families to have access to better services, wouldn't you agree?' Explain why this question is biased.
- 2 Rewrite the above question to eliminate bias.
- 3 A television program has a phone-in survey asking the question, 'Should criminals receive tougher sentences?' Explain why telephone phone-in surveys have a sampling bias.
- 4 Explain a better method of getting public opinion on this issue.

For each of the following, state if the data are categorical or quantitative. If categorical, state whether the data are nominal or ordinal, and if quantitative, state if they are continuous or discrete.

- 5 Customers in a video shop vote for their favourite movie.
- 6 Customers in a video shop have records kept on the number of movies they hire each year.
- 7 The video shop keeps records of the number of times each movie has been hired.
- 8 The video shop keeps records of the length of each movie.
- 9 Give an example of categorical data which is ordinal.
- 10 Give an example of quantitative data which is discrete.

Estimating populations

Statistics are used to estimate populations that are too difficult to count accurately. An example of this is the number of fish in a lake. The method used to estimate the population is the 'capture-recapture' technique.

In this technique, a sample of the population is captured, tagged and then released back into the population. A second sample is then captured. The percentage of the second sample that have been previously tagged (that is, recaptured) becomes an estimate of the percentage of the entire population that has been tagged.



Estimating a population

Make a large collection of marbles, jelly beans, matches or toothpicks.

- 1 Select a sample of these items and identify/tag them with a dab of liquid paper.
- 2 Put those selected back in with the others and mix them thoroughly.
- 3 Select another sample and see how many are tagged.
- 4 Estimate the total number.
- 5 Put those selected back and repeat the process to see if you get a similar answer.
- 6 Count the number of items to see how accurate your estimates were.

WORKED Example 8

A group of marine biologists want to estimate the fish population in a lake. To do this, they catch 100 fish, tag them and release them back into the lake. The next day, they catch 200 fish and find that 10 of them have been previously tagged. What would be their estimate of the population?

THINK

- 1 Calculate the percentage of the second sample that has been tagged.
- 2 Use this percentage to calculate what 100% of the population would be.
- 3 Give a written answer.

WRITE

$$\text{Percentage tagged} = \frac{10}{200} \times 100\% = 5\%$$

5% of population = 100
 so 1% of population = 20
 and 100% of population = 2000
 Their estimate of the fish population in the lake is 2000.

remember

1. The capture–recapture technique is used to estimate populations that cannot be accurately counted with ease.
2. A sample of the population is captured, tagged and released back into the population.
3. A second sample is then captured and the percentage tagged is an estimate of the percentage of the entire population that has been tagged.
4. The unitary method of percentages is then used to estimate the entire population.

EXERCISE 6E**Estimating populations****WORKED
Example****8**

- 1 The fish population of a river is to be estimated. A sample of 400 fish are caught, tagged and released. The next day another sample of 400 fish are caught and 40 of them have tags. Estimate the fish population of the river.
- 2 A colony of bats live near a school. Wildlife officers try to estimate the bat population by catching 60 bats and tagging them. These bats are then released and another 60 are caught, 9 of which had tags. Estimate the size of the bat population living near the school.
- 3 A river's fish population is to be estimated. On one day 1000 fish are caught, tagged and released. The next day another 1000 fish are caught. Estimate the population of the river if in the second sample of fish:
 - a 100 had tags
 - b 40 had tags
 - c 273 had tags.
- 4 A certain fish population is said to be endangered if the population falls below 15 000. A sample of 1000 fish are caught, tagged and released. The next day another sample of 1200 fish are caught, 60 of which had tags. Is the fish population endangered?
- 5 To estimate the fish population of a lake, 300 fish were caught. These 300 fish (150 trout, 100 bream and 50 perch) were tagged and released. A second sample of fish were then caught. Of 100 trout, 24 had tags; of 100 bream, 20 had tags; and of 100 perch, 8 had tags.
 - a Estimate the number of trout in the lake.
 - b Estimate the number of bream in the lake.
 - c Estimate the number of perch in the lake.
- 6 The kangaroo population in a national park is to be estimated. On one day, 100 kangaroos were caught and tagged before being released. (*Note:* For each sample taken, the kangaroos are released after the number with tags is counted.)
 - a The next day 100 were caught, 12 of which had tags. Estimate the population.
 - b The following day another estimate was done. This time 200 were caught and 20 had tags. Estimate the population again.
 - c A third estimate was done by catching 150 and this time 17 had tags. What will the third estimate for the population be?
 - d For a report, the average of the three estimates is taken. Calculate this average.



summary

Data collection

- A statistical investigation can be done by either census or sample.
- A census is when an entire population takes part in the investigation.
- A sample is when a small group takes part in the investigation and the results are taken to be representative of the whole group.
- There are three types of sample.
 1. Random sample — chance is the only factor in deciding who participates.
 2. Stratified sample — the sample taken is chosen so that it has the same characteristics as the whole population.
 3. Systematic sample — there is a method for deciding who participates in the sample.

Population characteristics

- A stratified sample can be used to ensure that the characteristics of your sample match the characteristics of the whole population.
- To select a stratified sample:
 1. Write each number in each stratum as a fraction of the whole population.
 2. Multiply by the size of the sample that you wish to take.
 3. Round off the answer for each stratum to the nearest whole number.
- Sometimes this method increases the sample size, but this may be necessary to get the best representation possible in your sample.

Bias

- Bias occurs when the results of an investigation are skewed to one side. This can occur because of:
 1. A poorly worded question that can lead the responder into a response favouring one side.
 2. A poorly chosen sample. Those participating in the investigation may not represent the whole population and be more inclined to a certain point of view.
 3. Misinterpretation of results. This can occur when a graph is drawn to give a certain impression.

Classification of data

- Data can be classified as being categorical or quantitative.
- Categorical data are data that are non-numerical. For example, a survey of car types is not numerical.
- Quantitative data are data that can be either counted or measured. For example, a survey of the daily temperature is quantitative.
- Quantitative data can be either discrete or continuous.
- Discrete data can take only certain values such as whole numbers.
- Continuous data can take any value within a certain range.

Estimating populations

Populations that can't be accurately counted with ease are estimated by using the capture–recapture technique.

CHAPTER

review

6A

6A

6A

6A

6B

6C

6D

- 1 For each of the following statistical investigations, state whether a census or a sample has been used.
 - a The average price of petrol in Sydney was estimated by averaging the price at 40 petrol stations.
 - b The Australian Bureau of Statistics has every household in Australia complete an information form once every five years.
 - c The performance of a cricketer is measured by looking at his performance in every match he has played.
 - d Public opinion on an issue is sought by a telephone poll of 2000 homes.

- 2 Name and describe three different methods for selecting a sample.

- 3 Which method of sampling has been used for each of the following?
 - a The quality-control department of a tyre manufacturing company road tests every 50th tyre that comes off the production line.
 - b To select the students to participate in a survey, a spreadsheet random number generator selects the roll numbers of 50 students.
 - c An equal number of men and women are chosen to participate in a survey on fashion.

- 4 Use a random number generator to select 10 numbers between 1 and 1000.

- 5 The table at right shows the number of students in each year of school.

In a survey of the school population, how many students from each year should be chosen, if a sample of 60 is selected using a stratified sample?

Year	No. of students
7	212
8	200
9	189
10	175
11	133
12	124

- 6 Bias can be introduced into statistics through:
 - a questionnaire design
 - b sample selection
 - c interpretation of statistical results.
 Discuss how bias could be a result of techniques in the above three areas.

- 7 State whether each of the following data types are categorical or quantitative.
 - a The television program that people watch at 7:00 pm
 - b The number of pets in each household
 - c The amount of water consumed by athletes in a marathon run
 - d The average distance that students live from school
 - e The mode of transport used between home and school

6D

- 8 For each of the quantitative data types below, determine if the data are discrete or continuous.
- a The dress sizes of Year 11 girls
 - b The volume of backyard swimming pools
 - c The amount of water used in households
 - d The number of viewers of a particular television program
 - e The amount of time Year 11 students spent studying

6E

- 9 To estimate the fish population of a lake, 100 fish are caught, tagged and released. The next day another 100 are caught and it is noted that 5 have tags. Estimate the population of the lake.

6E

- 10 Kimberley has a worm farm. To estimate the population of her farm, she catches 150 worms and tags them before releasing them. The next day, she catches 120 worms and finds that 24 of them have tags. Estimate the population of the worm farm.

6E

- 11 A sample of 200 fish are caught, tagged and released back into the population. Later Barry, Viet and Mustafa each catch a sample of fish.
Barry caught 40 fish and 3 had tags.
Viet caught 75 fish and 9 had tags.
Mustafa caught 55 fish and 7 had tags.
- a Find the estimate of the population that each would have calculated.
 - b Give an estimate for the population, based on all three samples.

Practice examination questions

1 multiple choice

Which of the following is an example of a census?

- A A newspaper conducts an opinion poll of 2000 people.
- B A product survey of 1000 homes to determine what brand of washing powder is used
- C Every 200th jar of Vegemite is tested to see if it is the correct mass.
- D A federal election

2 multiple choice

Which of the following is an example of a random sample?

- A The first 50 students to arrive at school take a survey.
- B Fifty students' names are drawn from a hat and those drawn take the survey.
- C Ten students from each year of the school are asked to complete a survey.
- D One class in the school is asked to complete the survey.



3 multiple choice

Which of the following is an example of categorical data?

- A** The type of car that is in each home
- B** The number of cars in each home
- C** The distance travelled by a person's car in a one year period
- D** The amount of money spent on petrol in a one year period

4 multiple choice

Which of the following is an example of continuous data?

- A** The type of car that is in each home
- B** The number of cars in each home
- C** The distance travelled by a person's car in a one year period
- D** The amount of money spent on petrol in a one year period

- 5** Carolyn is a marine biologist. She spends the day on a boat and 500 fish are netted. Carolyn notes the types of fish netted. There are 173 blackfish, 219 drummer and 108 mullet.

- a** Are Carolyn's data categorical or quantitative?
- b** The fish are tagged and released back into the school from which they were caught. Another 250 are then caught and it is noted that 63 have tags. Estimate the population of the school.

- 6 A paint company says that 1 litre of paint can paint an area of 4 m^2 . To test this, every 100th can is tested to see that it will cover at least 4 m^2 .
- a Are the data categorical or quantitative? If they are quantitative, are the data discrete or continuous?
 - b What type of sample has been used?
 - c The paint is mixed in one of five machines. Every can tested therefore comes from the same machine. A quality control officer then says that the data are biased. What is meant by the term bias?
 - d Explain what could be done to prevent the data from being biased.

